DOCUMENT RESUME

ED 173 805 CS 204 950

AUTHOR Lloyd-Jones, Richard
TITLE The Use of Tests of Writing and Skills Related to Writing.
PUB DATE Apr 79
NOTE 7p.; Paper presented at the Annual Meeting of the Conference on College Composition and Communication (30th, Minneapolis, Minnesota, April 5-7, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Composition (Literary); *Evaluation Criteria; Expository Writing; Higher Education; Student Evaluation; *Testing; Test Reliability; *Writing Skills; *Written Language

ABSTRACT

There are two modes for testing writing: atomistic and holistic. Atomistic tests depend on isolating a form, emphasizing the recognition of the form rather than the creation of it. Holistic tests require the generation of writing, so that the forms are seen in a particular context. The atomistic tests have great danger of leading to false inferences; their appropriate uses are limited, though valuable; and misuse is fairly common. Holistic tests come in two varieties, each of which focuses on the network which joins the separate elements of the writing: the Educational Testing Service's (ETS) test, which is designed to rank samples of writing in broad quality categories by means of tacit judgments, and the Primary Trait (PT) system, which describes papers in broad categories explicitly described in relation to purposes implied in the writing task. Problems exist in both systems in generating a sample of writing that is truly representative of the quality of writing done by the person being tested. As opposed to atomistic tests, holistic tests have difficulty in controlling the scoring for a large number of samples. The ETS system is probably more effective in ranking individual students, while PT provides more information for assessment and research. (DF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

ED173805

The Use of Tests of Writing and Skills Related to Writing

(CCCC, Minneapolis, 1979)

Richard Lloyd-Jones

In truth we can't test directly the ability to write, but in practice we have to approximate such tests, so we look at our budgets and decide what kinds of imperfect indicators offend us least as we try to satisfy our immediate needs. Such limits can be observed of all indirect testing, but the complexity of writing (both as a process and as a product) make the problems more nearly overwhelming, and the excessive claims based on the test results are more damaging to individuals because the act of writing is closely bound to one's sense of self within a social group.

The underlying difficulty is that, even though we usually want to know about writers, we are forced to examine specific pieces of writing acquired in ways almost guaranteed to inhibit performance--mere footprints from which we expect to detect the nature of a real person. When we examine writing in its worldly context, we may safely assume that it exists as a real message to carry out a real task, carrying an idea from the self to the other. One example may still be but a footprint of the whole person, but it is a print of a real step. In the test world such a footprint is artificially planted by telling a person to walk "naturally" from nowhere to nowhere as though she or he might not then place a foot awkwardly or indifferently. And even the most naive writing is much more difficult to examine than a mere footprint--especially if, as in most testing situations, we are examining the writing of many people gathered by the cartload.

Simply, we value a piece of writing because it serves its intended purpose, and purposes are specific to some situations. For example, a good set of directions will enable you to replace your furnace filter quickly, neatly, and safely. A good persuasive comment will prod you into seeing that the filter is changed. We are not obliged to ask how the writing worked if it did its task. If it fails, we may begin to talk of techniques, or forms, but we may also merely observe that the direction maker over-estimated the skill of his reader. That judgment is stated in terms of the purpose of explaining to a particular kind of person. Only when we begin to talk about the level of abstraction or the form of the verb or groupings of the sub-steps have we begun to deal in form.

Teaching and testing, thus, require some hypotheses about what elements in the writing accomplished the task. We need to be able to point at some element of the writing in order to say to the pupil that such-and-such is what made the writing work--go and do likewise. One trouble is that what works for explanation may inhibit persuasion--say, including directions to change clothes or wear an apron may make a person put off changing the filter--and what sells clean furnace filters (efficiency) may not sell a new hat (stylishness). Or, more relevantly here, the curtness of good directions may offend a person who needs to be led to a new idea. The relationships

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard Lloyd-Jones

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS204950

between purposes and forms are by no means clear even though experienced writers make pretty good guesses in familiar situations. A good technical writer may be inept at public relations euphemisms.

The issue is important here because tests deal in forms even though the world is concerned primarily with accomplished purposes. In general our tests are devised to make a writer show his or her command of a form which we are persuaded is important. A good test shows both the ability to construct the form and the judgment about the aptness of particular forms for a task. Even the cleverest among us make mistakes about the rightness of forms.

Tests of writing skills are either atomistic or holistic. The atomistic tests depend upon isolating a form, and usually emphasize recognition of the form rather than the creation of it. Most atomistic tests which try to represent contexts within "objective" test forms are essentially reading tests. Holistic tests require the generation of writing, so the forms are seen in a particular context and the papers are judged on their quality as a network of forms. In general it is easier to relate a network of forms to a purpose than it is to relate specific forms to purpose.

The atomistic tests are relatively cheap but very indirect, so the possibility of false inferences is great. We are asked to assume a lot of relationships. At one extreme are vocabularly tests--efforts to estimate the size and sophistication of a person's wordhoard. One assumes that a person has concepts and command of all language roughly proportional to one's vocabularly; for many purposes this seems adequate. We know anyway that vocabularly tests are useful predictors of many kinds of intellectual achievements, for example, success in college, although we also know that a number of well-worded people--professors for example--sometimes write badly. The correlation is far from perfect, but it is adequate to permit a rough sorting of people for some purposes, especially if individuals are given the options of retests or (more important) of presenting other kinds of evidence to show competence. The virtue of vocabulary tests is that they are cheap, and we know how to make them. We are even fairly competent at recognizing bias in the choice of particular items.

Atomistic tests of usage are more dangerous, perhaps partly because the choices being observed in the test items are more likely to be misinterpreted, more likely to relate to social dialect than to writing skill. Folks outside of school set great store by their own usage, especially in details they associate with school manners and firm discipline. That is, with prestigeous forms chosen by others. Syntactical choices which are fundamental to expressing relationships are generally so standard in the society that tests of such items do not provide a suitable distribution of scores; in order to make tests separate people, test items dealing with surface features of language in the process of change are most likely choices to provide a broad distribution. Hence the tested features are important more for social than intellectual reasons. Such tests are not necessarily bad predictors of success in college, for they point to a person who is familiar with the common values of the society, values which generally fit

well into the college environment. Such tests also eliminate some very good but unusual people. Still, we all know papers which are dull, pointless, and empty of information and which have no surface "errors." Once again the connection of details of usage with essential skill in writing may be very modest; by being somewhat plausible it often leads to considerable damage to students simply by being over-valued by people who are gullible about test scores.

More sophisticated features of language can be fit into the general standardized test patterns, and in general the more plausible tests are, the more hazard they may offer us. Still not all atomistic tests are cast in the standard multiple choice forms, interlinear re-drafting and cloze tests and sentence combining exercises can be used for tests. One may take even samples of prose, and (with or without the aid of computer memories) describe the sample in terms of the presence or absence of various predetermined forms--comma splices, periodic sentences, or even temporal organization. That is, one need not necessarily seek "errors", although in truth for most such testers "error" more than "rehtorical choice" is the common concern. We look for the negative, not the positive achievement. Even in those extreme forms of atomistic testing which depend on getting a sample of writing, the issue is that, although we ask a person to generate writing, we examine not the issue which holds it to some posited purpose, but pieces of language which can be observed independently of the purpose of the writer. The writing being described exists not as meaning but as a collection of forms. Often, it is simply a matter that we test what we know how to test whether or not it means anything.

For example, in all cycles of its testing the National Assessment of Educational Programs (NAEP) studied mechanics in the writing samples, and in the third cycle the assessment will even try to observe atomistically how writers handle various devices of cohesion. My colleague at Iowa, Paul Diehl, is currently developing a complex grid of linguistic features to permit comparison of details with overall effectiveness in handling different kinds of discourse. I doubt that such tests ought to be used to make judgments about individuals--and the National Assessment does not--but they are clearly useful in helping us discover something about the general relationship between effectiveness measured some other way--generally holistically--and the occurence of particular forms of language. Whether statistical control will be more informative than knowledge derived by centuries of tradition is still open to question, but such systematic observations may enable us to choose between competing bits of traditional advice.

You'll notice that I have not said that standardized or descriptive tests of atomistic elements have no use, but that the appropriate uses are limited, although valuable, and misuse is sufficiently common that I can symphathize with--on occasion even vote with--those who want blanket prohibitions against atomistic tests. I'll come back to that point.

Holistic tests come in two general varieties, both of which focus on the network which joins the separate elements of the writing. The more

commonly used and less expensive type is that associated with the Educational Testing Service. The more elaborate type is the Primary Trait or Thrust system developed for the National Assessment. In a tough sense the ETS system is designed to rank samples of writing in broad quality categories by means of tacit judgments. The Primary Trait system describes (and often values) papers in broad categories which are explicitly described in relation to purposes implied in the writing task. The descriptions are vaguely analogous to criteria in criterean referenced testing. Each of these systems has advantages and disadvantages. Neither is perfect for identifying good writers, but both come closer to dealing with writing than do any of the techniques dealing with the atoms.

Both systems have problems generating a sample of writing which is truly representative of the quality of writing done by the person being tested. A good writing task for test purposes appeals to the writer, is within his or her appropriate range of interests and needs, has a clearly defined purpose, does not irrelevantly favor one part of the test group over another, and can be addressed within the conditions of a testing situation. As much might be said of writing assignments in the classroom, but far more variation in response can be accepted in the classroom because evaluation is rarely based on only one score, students can search for ideas, instructors can help, and the physical environment for writing can be controlled to suit the writer. Some people write well ad lib, others require lots of incubation time Some write in the morning, some at night. Some do better with explanation, others prefer abstract exposition; still others prefer expressive narrative. Although creating a suitable task is a very troublesome problem for any tester, it perhaps is more serious for the Primary Trait tester, who must be especially careful to elicit one dominant rhetorical thrust or purpose for which the formal equivalents in language can be predicted. The more the limit is drawn precisely to one kind of siutation and purpose, the greater the problems of finding appropriate and stimulating tasks applicable for all of a given group of writers. Clearly Primary Trait tasks put great emphasis on imitating the purpose of real world discourse, and to some extent all PT scoring guides are couched in terms of satisfying the purpose. The underlying assumption is that people are not equally skilled in dealing with all kinds of writing.

Controlling the scoring for a large number of samples is also difficult. Atomistic tests may be turned over to electronic scorers, but writing samples require human scorers who are "calibrated" to report similarly about the papers they read. The tacit system of ETS is more easily managed, but in general less informative than the National Assessment's PT system. Essentially ETS tries to select experienced readers of student writing and then guides them experientially through the common reading of pre-scored samples in order to train the scorers to read similarly. Readings by more than one person are checked against each other in order to maintain consistency. Although in the training sessions some comments on the scoring of the sample papers are proper, still the intention of the calibration is to avoid having readers come to concentrate on any particular formal features. Such focus may distract the attention to details from the network. Tacit observations also leave readers flexible to make adjustments for papers which are idiosyncratic but none-the-less interesting and perhaps indicative of general writing talents. Good writing which does not exactly fit the

situation of the task may thus be honored. For predicting general achievements in writing or in giving credit or even in valuing a sample of writing for purposes of comparing the whole with some of the details, such a method is probably adequate and certainly easier to handle than PT.

For assessment purposes and for many research efforts one may want more explicit information about what is exemplified in writing marked with different scores. It is not enough to say merely that paper 1784 is in the upper quartile of 5000 papers. The PT descriptions may be either neutral or evaluative, although given the needs of most situations where tests are used, most are evaluative. Each score point is elaborately described both in examples of the formal features expected in relation to success in performing a particular task and in the level of success expected by the combination. Experienced writers can ordinarily predict the appropriate features quite well, and thus can create a suitable guide in advance, although a check against what is produced by writers in a pre-test sample or tryout of the task undoubtedly is useful. It is hard to devise a task which some bright person will not address successfully in unexpected ways and, anyway, even good versatile writers have limits on their experience. At the moment we lack systematic descriptions of how details of form relate to purpose, so we must depend on experience. Overall, the guide must stress purpose, because a problem caused by the formal explicitness of the guides is that some scorers are tempted to think a paper is no more than the sum of the named parts. The scoring guide must remind the readers that they are looking for the success in handling the primary rhetorical problem, not in presenting forms. Since the system is not likely to be used in placing or exempting students, but rather in describing or guiding the work of groups, one can safely penalize "good" writing which does not address the assigned task, although necessarily such procedures make the group seem less competent than it is. Still, the goal of the test is to recognize writing based on appropriate strategies, and we generally posit that performance is lower than competence.

The PT scores, however much they may undervalue writers, imply all of the information in the scoring guides and thus suggest the limits of interpretation as part of the report. Any report of test scores can be overvalued and usually is, but PT reports have some built-in protection because so much space must be devoted to explaining the guide within the situation of the task. Also because the PT score makes explicit the criterea of which the descriptions are based, they permit more kinds of correlation with various sub-scores. Given the present state of our ignorance about writing taken in large collections instead of one-at-a-time, this makes the method useful when researchers try to avoid errors of false association. In general the persuasiveness of the score depends less on credentials of the readers than on the plausibility of the scoring guide, so somewhat less expensive readers can be used, but the system is still harder to use and usually more costly than ETS holistic. Bad tasks are likely to be more visibly bad.

How, then, do we choose tests?

Atomistic tests are not tests of writing, but may relate to knowledge sufficiently like writing skill that such tests can be used for prediction--provided one gives the writers chances to challenge the test results and

simple minded interpreters are kept away from the scores.

Writing samples are still indirect tests of the ability to manage apprpriate language forms. But at least they demonstrate the value of the network, of the interaction of forms.

The ETS system probably works better in ranking individual students; PT provides more information for assessment and research.

In the end, though, we are not talking about any system with perfect results. We are dealing with the best possible approximations, so any statement about the quality of writing as measured outside of real contexts should be carefully limited. The "score" is merely a hypothesis to be checked against other observations.

Richard Lloyd-Jones
The University of Iowa